Making LLMs Better Many-to-Many Speech-to-Text Translators with Curriculum Learning

Yexing Du[1,2] **Youcheng Pan**[2] **Ziyang Ma**[3] **Bo Yang**[2] **Yifan Yang**[3]
Keqi Deng[4] **Xie Chen**[3] **Yang Xiang**[2*] **Ming Liu**[1,2*] **Bing Qin**[1,2]
[1]Harbin Institute of Technology [2]Pengcheng Laboratory
[3]Shanghai Jiao Tong University [4]University of Cambridge
{yxdu, mliu}@ir.hit.edu.cn, {panych,xiangy}@pcl.ac.cn

Abstract

Multimodal Large Language Models (MLLMs) have achieved significant success in Speech-to-Text Translation (S2TT) tasks. While most existing research has focused on English-centric translation directions, the exploration of many-to-many translation is still limited by the scarcity of parallel data. To address this, we propose a three-stage curriculum learning strategy that leverages the machine translation capabilities of large language models and adapts them to S2TT tasks, enabling effective learning in low-resource settings. We trained MLLMs with varying parameter sizes (3B, 7B, and 32B) and evaluated the proposed strategy using the FLEURS and CoVoST-2 datasets. Experimental results show that the proposed strategy achieves state-of-the-art average performance in 15×14 language pairs, requiring fewer than 10 hours of speech data per language to achieve competitive results. [1]

1 Introduction

Speech-to-Text Translation (S2TT) involves converting speech from a source language into text in a target language. Traditionally, S2TT tasks have relied on a cascaded system, as shown in Figure 1(a), where an Automatic Speech Recognition (ASR) module transcribes speech into text (Baevski et al., 2020; Gulati et al., 2020), followed by a Machine Translation (MT) module that translates the transcribed text into the target language (Cheng et al., 2019; Beck et al., 2019). However, this cascade system often suffers from error propagation (Sperber and Paulik, 2020). Recently, Multimodal Large Language Models (MLLMs), illustrated in Figure 1(b), have demonstrated advantages in simplifying model architecture and mitigating error propagation in both ASR (Zhang et al., 2023; Ma et al., 2024) and S2TT tasks (Chu et al., 2024).

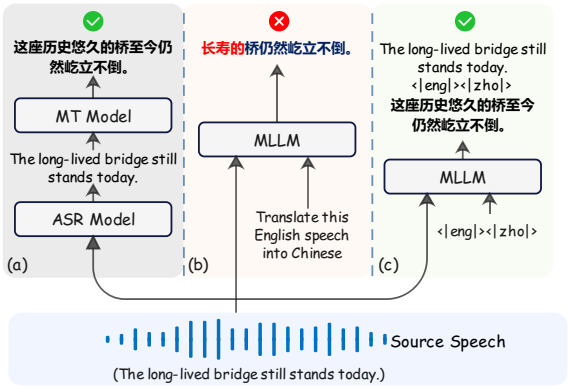

Figure 1: **Comparison of S2TT Methods.** (a) adopts a cascaded system; (b) directly generates translated text; (c) generates both transcription and translation text in an end-to-end process, with <|eng|><|zho|> indicating transcribing English and translating it into Chinese.

Current MLLMs process {*speech*, *instruction*} inputs to directly generate {*translation*}, but this approach heavily relies on large-scale S2TT datasets. Existing datasets (Wang et al., 2020c; Di Gangi et al., 2019) predominantly focus on English, while datasets supporting many-to-many S2TT, such as FLEURS (Conneau et al., 2022), remain limited. Meanwhile, Large Language Models (LLMs) have demonstrated strong many-to-many multilingual MT capabilities. This raises the question: can the MT capabilities of LLMs be effectively transferred to the S2TT task with limited data?

Inspired by advances in transfer learning (Pham et al., 2024; Mueller et al., 2024), we transform the S2TT task into a **Speech Recognition and Translation (SRT)** task, which involves training {*speech*, *instruction*} to generate {*transcription*, *translation*}, as shown in Figure 1(c). LLMs possess robust MT capabilities, which can be adapted to S2TT tasks with minimal data. This approach allows MLLMs to harness MT capabilities for many-to-many S2TT, effectively combining the advantages of both cascade and end-to-end models.

*Corresponding author.

[1]The source code and models are released at https://github.com/yxduir/LLM-SRT.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12466–12478
July 27 - August 1, 2025

To connect MT and S2TT tasks, we propose a three-stage curriculum learning strategy: (1) **ASR**, which trains the MLLM for multimodal alignment, enabling the model to understand speech and generate transcriptions; (2) **Speech-Aided Machine Translation (SMT)**, where both speech and transcription are provided, and the MLLM generates translations to improve cross-lingual capabilities; (3) **SRT**, where only speech is provided, and the MLLM generates both transcription and translation. The training proceeds sequentially through these three stages, with each stage resuming from the checkpoint of the previous one. The resulting MLLM achieves many-to-many S2TT through the SRT task. Additionally, we designed specialized instructions for many-to-many S2TT and implemented an optimized lightweight speech adapter for efficient speech feature compression to accelerate inference.

To evaluate our strategy, we trained three MLLM variants (3B, 7B, and 32B). In low-resource scenarios (with fewer than 10 hours of data per language), our MLLM, trained on the FLEURS dataset, demonstrated strong many-to-many S2TT capabilities, outperforming existing state-of-the-art end-to-end models. We also assessed performance on the EN-X direction of the CoVoST-2 dataset, where sufficient training data is available, and found that our strategy remains effective, surpassing state-of-the-art models.

The key contributions of this work are:

- This paper adopts a strategy that transforms the S2TT task into an SRT task, leveraging the machine translation capabilities of LLMs to enhance the many-to-many S2TT performance of MLLMs, particularly in low-resource settings.
- We propose a three-stage curriculum learning strategy and systematically evaluate our strategy across datasets of varying scales and model sizes (3B, 7B, 32B). To the best of our knowledge, our model is the first MLLM to support many-to-many S2TT at the 32B scale.
- Our model achieves state-of-the-art average performance across 15×14 translation directions in the S2TT task under low-resource settings on the FLEURS dataset, while also demonstrating strong robustness on the CoVoST-2 dataset in high-resource scenarios.

2 Methodology

In this section, we present the methodology of our approach. Section 2.1 defines the tasks involved in our method. Section 2.2 presents the architecture of the LLM-SRT model. Section 2.3 explains our curriculum learning strategy, which sequentially fine-tunes the model for ASR, SMT, and SRT tasks.

2.1 Problem Formulation

In this section, we define the following tasks:

ASR: Given the audio input $\mathbf{X}$ and the instruction text $\mathbf{T}$, the goal is to produce the transcribed text $\mathbf{Y}$.

SMT: Given the audio input $\mathbf{X}$, its corresponding transcription $\mathbf{Y}$, and the instruction text $\mathbf{T}$, the goal is to produce the translated text $\mathbf{Z}$.

SRT: Given the audio input $\mathbf{X}$ and the instruction text $\mathbf{T}$, the goal is to produce both the transcription $\mathbf{Y}$ and the translated text $\mathbf{Z}$.

2.2 LLM-SRT

The LLM-SRT architecture is shown in Figure 2. The speech encoder extracts features from the speech input, and the speech adapter layer connects these features to the LLM, aligning their dimensions and incorporating speech feature compression. Finally, the LLM generates textual output by processing the concatenated embeddings derived from both text and speech features.

Speech Encoder. The speech encoder processes the audio input $\mathbf{X}$ into a high-dimensional representation using the frozen Whisper encoder (Radford et al., 2023), which has been pretrained on large-scale supervised datasets for speech recognition and translation.

$$\mathbf{H} = \text{Encoder}(\mathbf{X}), \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{T \times D}$ is the encoder's output, with T representing the time dimension and D the hidden dimension of the encoder.

Speech Adapter. The speech adapter compresses the time dimension T and adjusts the hidden dimension D to match the LLM's hidden dimension d_{LLM}.

We use a Q-Former to convert input sequences into fixed-length query representations.

$$\mathbf{Q}' = \text{Q-Former}(\mathbf{Q}, \mathbf{H}), \tag{2}$$

where $\mathbf{Q}' \in \mathbb{R}^{n_q \times D_q}$ is the output of the Q-Former, $\mathbf{Q} \in \mathbb{R}^{n_q \times D_q}$ is the trainable query matrix, n_q is the

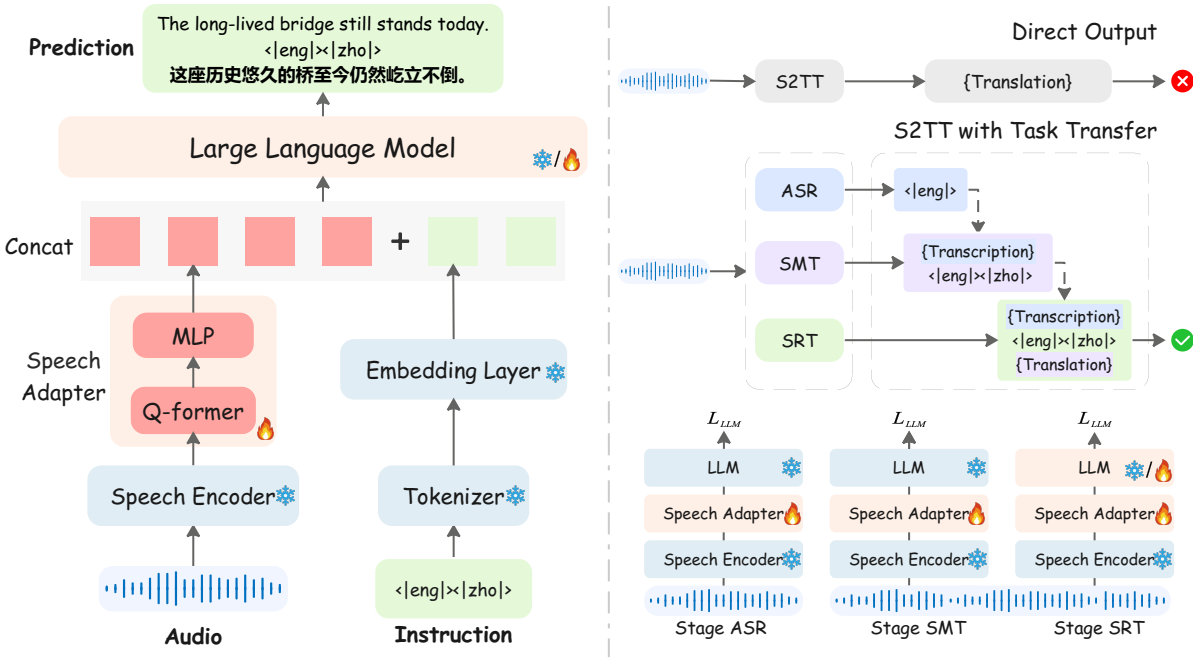

Figure 2: **The Architecture of LLM-SRT.** LLM-SRT consists of a speech encoder, speech adapter, and LLM. A three-stage curriculum learning strategy sequentially trains the ASR, SMT, and SRT tasks, as shown in Table 1. In stages 1 and 2, the speech adapter is continuously trained to enable efficient fine-tuning. In stage 3, the LLM is additionally unfrozen, while the speech adapter continues to be trained.

Task	Audio	Instruction	Prediction
ASR	✓	<|eng|>	Will it rain tomorrow?
ASR	✓	<|zho|>	明天会下雨吗?
SMT	✓	Will it rain tomorrow?<|eng|><|deu|>	Regnet es morgen?
SMT	✓	明天会下雨吗? <|zho|><|jpn|>	明日は、雨かな?
SRT	✓	<|eng|><|deu|>	Will it rain tomorrow?<|eng|><|deu|>Regnet es morgen?
SRT	✓	<|zho|><|jpn|>	明天会下雨吗? <|zho|><|jpn|>明日は、雨かな?

Table 1: **Instruction Design.** The instruction design is intended for fine-tuning instructions for three tasks: ASR, SMT, and SRT, using simple yet effective instructions to distinguish between them.

number of queries, and D_q is the hidden dimension of the Q-Former.

After the Q-Former layer, a multilayer perceptron (MLP) projects the feature dimensions from D_q to d_{LLM}:

$$\mathbf{E}^{\mathbf{X}} = \text{ReLU}(\mathbf{Q}'\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \quad (3)$$

where $\mathbf{E}^{\mathbf{X}} \in \mathbb{R}^{n_q \times d_{\text{LLM}}}$ is the output of the MLP layer, ready for LLM processing.

Tokenizer and LLM. The tokenizer and embedding layer process the instruction $\mathbf{T}$ and produce $\mathbf{E}^{\mathbf{T}} \in \mathbb{R}^{n_t \times d_{\text{LLM}}}$, where n_t is the length of the text tokens.

The speech features $\mathbf{E}^{\mathbf{X}}$ and text features $\mathbf{E}^{\mathbf{T}}$ are concatenated and fed into the frozen LLM:

$$\mathbf{E}^{\mathbf{Z}} = \mathbf{E}^{\mathbf{X}} \oplus \mathbf{E}^{\mathbf{T}}, \quad (4)$$

where $\mathbf{E}^{\mathbf{Z}} \in \mathbb{R}^{(n_t+n_q) \times d_{\text{LLM}}}$ is processed by the LLM to generate the output text.

The output text varies depending on the task. During training, the parameters of the adapter layer are updated based on the loss of the LLM's output.

2.3 Curriculum Learning

LLM-SRT adopts a curriculum learning approach that incorporates three training tasks: ASR, SMT, and SRT.

Instruction Design. We designed minimalist instructions to help the model distinguish between tasks while reducing the instruction token length, as shown in Table 1. This design ensures that instructions like <|eng|><|deu|> appear in the generated answers, effectively segmenting transcription and translation content in the SRT task.

ASR. In this stage, the model is pre-trained to develop ASR capabilities with a focus on multimodal alignment, while expanding language support by training on all intended languages. The speech adapter is trained with as much data as possible to ensure efficient fine-tuning.

SMT. This stage enhances the model's cross-lingual abilities. Starting from the ASR checkpoint, the model takes both transcribed text and audio as input to generate translations based on the instruction. The purpose of this step is to activate the LLM's inherent machine translation capabilities and establish the connection between the MT and S2TT tasks.

SRT. This stage activates the SRT capabilities of the MLLM, finalizing the model. Training continues from the SMT checkpoint, with the model receiving only audio input and a task-specific instruction, outputting both the transcription and translation of the speech. This extends the MT capabilities of LLMs to the S2TT task.

3 Experiments

3.1 Datasets

FLEURS. FLEURS[2] (Conneau et al., 2022) serves as the speech counterpart to the FLoRes[3] (Team et al., 2022) machine translation benchmarks. It includes 102 languages, with each training set containing approximately 10 hours of supervised speech data per language.

CoVoST-2. CoVoST-2[4] (Wang et al., 2020c) is a large-scale multilingual S2TT corpus derived from the Common Voice dataset (Ardila et al., 2020). It contains translations from 21 languages to English and from English to 15 other languages.

3.2 Experiment Settings

Model Architecture. The baseline model consists of an LLM (Qwen 3B, 7B, 32B), a frozen speech encoder (Whisper-large-v3), and a trainable adapter layer comprising a Q-Former and an MLP. Following the configuration in Yu et al. (2024), we use 80 queries, each with a dimension of 768. Training can be minimized by freezing the LLM, or LoRA (Hu et al., 2021) can be applied for training.

Training Details. We used bf16 precision with Distributed Data Parallel (DDP), a learning rate of 1×10^{-4}, 1000 warmup steps, and the AdamW optimizer. The models were trained on four A100 GPUs. We provide detailed settings in Table 11, 12, 13, and 14 of the Appendix.

For the ASR task, we used the Common Voice[5] dataset, and for the SMT and SRT tasks, we used the FLEURS and CoVoST-2 datasets. In a 4-card A100 environment, the 3B and 7B models train in 3 days, while the 32B model trains in 7 days.

3.3 Compared Methods

We compare both cascade and end-to-end S2TT models, all of which support many-to-many S2TT.

- **Cascaded systems** are pipeline-based approaches, where an ASR model first transcribes speech into text, which is then translated by a machine translation model.
- **SeamlessM4T**[6] (Barrault et al., 2023) is a foundational multilingual and multitask model capable of seamlessly translating and transcribing both speech and text. It supports speech-to-text translation for nearly 100 input and output languages.
- **Qwen-Audio**[7] (Chu et al., 2023) is the multimodal extension of Qwen-LLM, designed to process diverse audio modalities, including speech, natural sounds, music, and songs, alongside text, generating text-based outputs.

Evaluation Metric. We use WER[8] (Morris et al., 2004)(for the ASR task) and BLEU[9] (Post, 2018) (for the S2TT task) as evaluation metrics.

Metric	Details
WER	Text normalization follows Whisper
BLEU	SacreBLEU signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3 Except for jpn, kor, tha, yue, zho with char: nrefs:1|case:mixed|eff:no|tok:char|smooth:exp|version:2.4.3

Table 2: **Metric Details.** We followed the settings of SeamlessM4T-V2 (Barrault et al., 2023).

[2] https://huggingface.co/datasets/google/fleurs

[3] https://huggingface.co/datasets/facebook/flores

[4] https://github.com/facebookresearch/covost

[5] https://commonvoice.mozilla.org/en/datasets

[6] https://github.com/facebookresearch/seamless_communication

[7] https://github.com/QwenLM/Qwen-Audio

[8] https://huggingface.co/spaces/evaluate-metric/wer

[9] https://github.com/mjpost/sacrebleu

X→12 Languages	S2TT Data (hour)	FLEURS Eng	Deu	Fra	Jpn	Rus	Zho	Avg.
Cascaded ASR+MT Methods								
Whisper + Qwen2.5-3B	-	23.4	21.0	20.9	13.6	19.6	14.5	18.8
Whisper + Qwen2.5-7B	-	26.7	23.3	22.5	15.1	21.5	16.4	20.9
Whisper + Qwen2.5-32B	-	29.9	26.0	24.6	17.3	23.8	18.5	23.3
End-to-End Models								
SeamlessM4T-V2 (2.3B)	351,000	33.1	20.5	19.6	13.2	19.6	15.2	20.2
Qwen2-Audio (7B)	in-house	22.6	20.1	20.6	4.0	15.0	13.7	16.0
Baseline-3B	52	11.8	9.0	9.5	5.2	9.7	6.2	8.6
LLM-SRT-3B	52	27.2	**22.6**	**22.0**	**14.3**	**21.3**	**16.5**	**20.6**
LLM-SRT-7B	52	27.4	**23.7**	**22.8**	**15.5**	**21.8**	**16.9**	**21.4**
LLM-SRT-32B	52	32.5	**26.8**	**26.1**	**17.5**	**25.6**	**19.2**	**24.6**

Table 3: **BLEU Scores on 6x12 Directions in FLEURS.** Underlined denotes previous state-of-the-art end-to-end models, while **bold** indicates models that outperform them. "-" indicates no S2TT data was used due to the cascade system. The baseline uses the same instruction-tuning strategy as Qwen2-Audio.

3.4 Overall Results

As shown in Tables 3 and 4, we evaluate the S2TT performance in low-resource settings on the FLEURS dataset. The results indicate that our model achieves state-of-the-art performance in the many-to-many S2TT task. Similarly, Table 5 presents the results under high-resource conditions on the CoVoST-2 dataset. Table 7 provides a comparison of inference speed, and Table 8 presents an ablation study of the three-stage curriculum learning.

Language Support. As an LLM-based model designed for many-to-many S2TT, conducting comprehensive baseline comparisons is both essential and challenging. To ensure a thorough evaluation, we compare baselines across 6×12 language directions and benchmark our model against state-of-the-art approaches in the 15×14 setting. The supported languages are listed in Table 10, while the complete experimental results are provided in Table 18.

Baseline-3B vs. LLM-SRT-3B. For the baseline model, we first conduct ASR pretraining as in Qwen2-Audio, followed by S2TT instruction-tuning with the same setup. Due to limited data, the baseline performs poorly, highlighting the limitations of traditional fine-tuning in low-resource settings.

In contrast, our curriculum learning training strategy achieves state-of-the-art performance (8.6→20.6) in low-resource scenarios, as we transform S2TT into an SRT task, effectively leveraging the machine translation capability of the LLM to achieve many-to-many S2TT.

SeamlessM4T-V2 vs. LLM-SRT-3B. The LLM-SRT-3B demonstrates superior performance over SeamlessM4T-V2 on non-English languages (e.g., for French 18.8→22.0), while it lags behind SeamlessM4T-V2 in English-to-X translation. This discrepancy can largely be attributed to the larger amount of S2TT data available for SeamlessM4T-V2, which includes 351,000 hours of training data compared to just 52 hours for LLM-SRT-3B. In situations where data resources are limited, our LLM-based approach has a greater advantage.

Cascaded Systems vs. LLM-SRT. As shown in Table 3, when using the same LLM as in the cascaded system, our MLLM demonstrates a clear performance advantage (e.g., for 3B, 18.8→20.6), highlighting the benefits of the end-to-end approach. The MLLM's superior performance stems from its integrated framework, which eliminates intermediate steps and enables more efficient knowledge transfer, resulting in improved translation accuracy and robustness. This makes the model more effective in handling complex language tasks.

Scaling Law of LLM-SRT. As shown in Table 3, experiments on models with 3B, 7B, and 32B parameters demonstrate that our method follows the scaling law of LLMs (e.g., 20.6 for 3B, 21.4 for 7B, and 24.6 for 32B). Notably, the 32B model achieved state-of-the-art performance across all directions, confirming the generalizability of our approach. Our model's performance is strongly correlated with the machine translation capability of the LLM, making the choice of an appropriate LLM foundation crucial for optimal performance.

X→14 Languages	S2TT Data (hour)	FLEURS Eng	Deu	Fra	Ind	Ita	Jpn	Kor	Nld	Por	Rus	Spa	Tha	Vie	Yue	Zho	Avg.
Machine Translation																	
Qwen2.5-3B	-	29.5	25.4	25.2	24.6	22.7	18.4	18.8	22.1	27.1	23.4	22.1	18.3	22.6	19.0	20.5	22.6
Speech to Text Translation																	
SeamlessM4T-V2 (2.3B)	351,000	33.3	21.6	21.1	18.4	19.1	14.5	17.2	18.4	18.8	20.6	17.6	12.8	17.5	15.2	16.7	18.8
SeamlessM4T-V2 +Lora	351,129	32.8	22.0	21.3	19.1	19.4	14.4	17.0	18.6	18.6	21.3	18.1	13.2	17.5	14.9	17.0	19.0
LLM-SRT-3B-V2	129	27.8	23.4	24.0	22.4	22.0	15.3	17.9	20.4	25.7	22.8	21.9	12.9	18.2	15.5	19.2	20.6
LLM-SRT-3B-V2 +Lora	129	29.1	**24.5**	**24.3**	**22.9**	**22.8**	**16.5**	**18.0**	**20.9**	**26.2**	**23.3**	**22.6**	**14.7**	**19.0**	**16.3**	**19.6**	**21.4**

Table 4: **BLEU Scores on 15×14 Directions.** The complete results are in Table 15 and 18 in the Appendix.

Many-to-Many S2TT on FLEURS. As shown in Tables 4 and 18, we compared performance across 15 languages and 210 translation directions. Table 4 reports the average performance across the 15 languages, where our model achieves state-of-the-art BLEU performance (18.8→21.4). Table 11 presents detailed results for all 210 directions, showing that our model outperforms SeamlessM4T-V2 in 154 directions.

Train Adapter Only vs. Fine-tune LLM. As shown in Table 4, LLM-SRT-3B-V2 achieves high translation performance (20.6) by freezing the speech encoder and LLM, while training only the speech adapter. Further performance improvement (20.6→21.4) can be achieved by unfreezing the LLM, such as through LoRA training.

MT vs. S2TT. As shown in Figure 3, we compared the MT performance of Qwen2.5-3B with the S2TT performance of LLM-SRT-3B-V2 across 210 translation directions. The results show a strong correlation between our MLLM's S2TT and MT performance, confirming that the strong S2TT capability of LLM-SRT-3B stems from the LLM's machine translation ability.

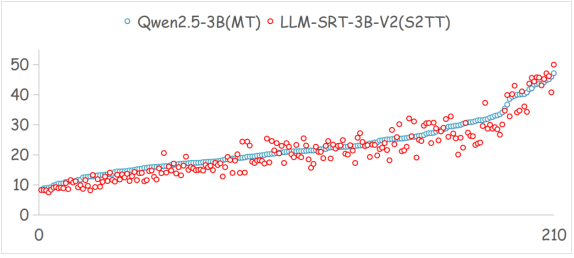

Figure 3: **BLEU Scores for 15×14 Directions: Comparison between MT and S2TT.** The results show a strong correlation, suggesting that our S2TT capability is derived from the MT model. Table 16 includes an error analysis showing that S2TT outperforms MT.

Eng→X	S2TT Data (hour)	CoVoST-2 Deu	Jpn	Zho	Avg.
Cascaded ASR+MT Methods					
Whisper+NLLB-3.3B	-	33.4	31.0	32.0	32.1
Whisper+Qwen2.5-3B	-	21.1	28.1	35.0	28.0
Whisper+Qwen2.5-7B	-	24.1	29.8	37.6	30.5
Whisper+Qwen2.5-32B	-	28.3	35.0	41.1	34.8
End-to-End Models					
SeamlessM4T-V2	351,000	37.0	39.7	35.9	37.5
Qwen-Audio (7B)	3,700	25.1	-	40.3	-
Qwen2-Audio (7B)	in-house	29.9	39.7	44.0	37.9
with FLEURS Data					
LLM-SRT-3B	52	24.9	37.3	40.7	34.3
LLM-SRT-7B	52	26.6	40.1	41.4	36.0
LLM-SRT-32B	52	30.2	**43.1**	43.5	**38.9**
with CoVoST-2 Data					
Baseline-3B*	430	23.6	34.7	38.6	32.3
Baseline-7B*	430	25.3	36.5	40.2	34.1
LLM-SRT-3B*	430	26.5	39.4	44.0	36.6
LLM-SRT-7B*	430	28.7	**41.6**	**47.1**	**39.1**

Table 5: **BLEU Scores on CoVoST-2.** * indicates trained on CoVoST-2 dataset.

Eng→X S2TT. As shown in Table 4, the relatively weaker performance of our model in Eng→X translations is primarily due to the limited amount of English data, stemming from the balanced FLEURS dataset, which contains fewer than 10 hours of data per language. Consequently, as shown in Table 5, our 3B model underperforms compared to SeamlessM4T-V2, but our method is more efficient in data-constrained scenarios.

Scaling Law of S2TT Data. As shown in Table 5, we compare the performance of Eng→X translations in both low-resource and high-resource scenarios, with data scales ranging from 52 hours to 430 hours. The results indicate that both our 3B (34.3→36.6) and 7B (35.0→39.1) models show consistent performance improvements as the data scale increases, demonstrating that our approach is effective in low-resource settings and scales well with more data.

MLLM Architecture. Our Baseline-7B model follows the Qwen2-Audio setup but keeps the speech encoder frozen, leading to lower performance (34.1 vs. 37.9).

LLM-SRT-7B employs a curriculum learning strategy that incrementally fine-tunes only the adapter layer, allowing the model to adapt more effectively while maintaining efficiency. This approach leads to notable improvements in performance by enabling stable and gradual optimization. Although unfreezing both the speech encoder and the LLM could yield further performance gains, it would also result in significantly higher computational costs.

Multi-task Model LLM-SRT. LLM-SRT is a multi-task model that supports ASR, SMT, and SRT tasks. As shown in Table 6, we evaluated the performance of all tasks.

We found that performing the SRT task did not degrade, but rather slightly improved, ASR performance. Moreover, with the correct transcription, the model achieved a high BLEU score in the SMT task.

Model	Task	WER↓	Deu	Jpn	Zho
LLM-SRT-7B* Eng→X	ASR	11.1	-	-	-
	SMT	0	32.8	43.6	55.6
	SRT	10.9	28.7	41.6	47.1

Table 6: **Performance of Different Tasks.** We evaluated the model on the CoVoST-2 dataset and found that the SMT task, which uses both speech and ground-truth transcription as inputs, achieved a notably high BLEU score.

Inference Speed. As shown in Table 7, our method achieves nearly a 3x improvement in inference speed compared to the Qwen2-Audio model, which has a similar parameter size. Even with beam search enabled (5 beams) in our model, the speed remains faster than the greedy search of Qwen2-Audio.

The speed difference is mainly due to our optimized speech adapter design, which compresses the audio features to a fixed size of 80, significantly reducing the token length input to the LLM and accelerating inference.

3.5 Ablation Study

Effect of the ASR, SMT, and SRT Tasks. Initially, ASR training was performed to establish a strong baseline (Bansal et al., 2018; Le et al., 2023). Removing ASR pre-training resulted in a 2.3-point decrease in BLEU score, emphasizing its importance in S2TT. This section explores the effect of skipping the SMT task and proceeding directly to SRT. As shown in Table 8, removing SMT and SRT resulted in performance drops of 1.1 and 4.9 points, respectively. While direct SRT maintains an MT-S2TT link with only minor degradation, omitting SRT and relying solely on instruction fine-tuning leads to a substantial performance drop.

Model	Strategy	Batch	Time(s)↓
Qwen2-Audio	Greedy Search	4	59
		8	/
	Beam Search 5	4	/
LLM-SRT-7B*	Greedy Search	4	74
		8	39
		16	28
		32	22
		64	**19**
	Beam Search 5	4	93
		8	64
		12	56

Table 7: **Inference Speed Comparison.** We compared the inference time for processing 1,000 speech samples between Qwen2-Audio and LLM-SRT-7B, both with similar parameter sizes, using a 4-card 4090 DDP FP16 inference setup. LLM-SRT-7B demonstrated a 3x speed improvement. / Indicates out-of-memory issues.

Model	Deu	Jpn	Zho	Avg.
LLM-SRT-7B*	28.7	41.6	47.1	39.1
w/o ASR	26.4(-2.3)	38.6(-3.0)	45.5(-1.6)	36.8(-2.3)
w/o SMT	27.6(-1.1)	39.7(-1.9)	46.5(-0.6)	38.0(-1.1)
w/o SRT	25.6(-3.1)	36.7(-4.9)	40.4(-6.7)	34.2(-4.9)

Table 8: **Ablation Study.** We evaluated the model on the CoVoST-2 Eng→X dataset. Removing the SRT task leads to a substantial reduction in BLEU score, underscoring its critical role in overall performance.

Case Study. As shown in Tables 9 and 17, SeamlessM4T-V2 demonstrates poor performance in Japanese, achieving the lowest BLEU score. Qwen-Audio outperforms it, while our method outperforms significantly. Our approach follows a two-step process: first, it generates a transcription of the input speech, which is then used to produce the translation. This strategy ensures that the translation benefits from the transcribed text, leveraging its structure and context to improve accuracy. By incorporating transcription into the translation process, our method minimizes ambiguities and enhances translation quality, especially in complex or context-dependent scenarios.

Model		BLEU↑
Audio	三国是中国古代历史上最血腥的时代之一。成千上万的人为了争夺西安豪华宫殿最高的权力而死去。	
Ground-truth	三国志は、古代中国の歴史の中で最も血なまぐさい時代の1つでした。西安の大宮殿の最高位を狙う争いの中で何千人もが戦士しました。	
SeamlessM4T-V2	三国は中国古代歴史で最も血腥な時代の一つ千上万の人がシ アン豪華宮殿の最高権力を争うために死ぬ huan	14.7 | *5.6*
Qwen2-Audio	三国は中国の歴史で最も血なましい時代の一つです。何千人もの人が、西安の豪華宮殿の権力を得るために死んでいます。	27.9 | *12.5*
LLM-SRT-3B	三国是中国古代历史上最血腥的时期之一，成千上万人为了争夺西安豪华宫殿的最高权力而死去。<|zho|><|jpn|>	
	三国時代は中国の歴史上で最も血なまぐさい時代の一つで、西安の豪華な宮殿で最高の権力を争うために、数万人が死にました。	34.9 | *16.5*
LLM-SRT-32B	三国是中国古代历史上最血腥的时代之一。成千上万的人为了争夺西安豪华宫殿的最高权力而死去。<|zho|><|jpn|>	
	三国時代は、中国の歴史の中で最も血なまぐさい時代の1つで、西安の豪華な宮殿の支配権を巡って何千人もの人々が死にました。	49.3 | *40.4*

Table 9: **Case Study.** We compare the BLEU scores of our method with those of other approaches, using the 'char' tokenizer (denoted in regular font) and the 'ja-mecab' tokenizer (presented in *italics*). With a transcription Character Error Rate (CER) score of 11.4 for the 3B model and 4.6 for the 32B model.

4 Related Work

Cascaded S2TT. This method follows a two-step process: first, ASR transcribes the spoken language into text, and then MT translates the transcribed text into the target language. This approach leverages the strengths of specialized ASR (Radford et al., 2023; Baevski et al., 2020) and MT (Fan et al., 2021) models, utilizing extensive training data and advanced techniques. ASR models accurately convert speech to text, while sophisticated MT models, benefiting from large multilingual datasets, translate with high accuracy and fluency. However, the cascaded approach is prone to error propagation.

End-to-End S2TT. In this paradigm, a single model is trained to directly map speech from the source language to text in the target language, skipping the intermediate transcription step (Wang et al., 2020a,b; Inaguma et al., 2020). Early work on joint speech recognition and translation primarily used streaming models, aiming to provide real-time multilingual synchronization (Sperber et al., 2020; Dong et al., 2021; Papi et al., 2024). These pioneering efforts focused more on reducing latency and enhancing efficiency than offline speech translation systems. End-to-end ST offers several advantages, including reduced latency, simplified system architecture, and the elimination of error propagation between the ASR and MT stages. Despite these benefits, end-to-end ST models face challenges, such as the need for extensive parallel speech-to-text data, which is resource-intensive and difficult to obtain.

Audio MLLMs. Recent advancements in audio MLLMs (Li et al., 2025) have significantly improved speech recognition and translation. SpeechGPT (Zhang et al., 2023) uses prompting to enhance speech recognition in large language models. BLSP-KD (Wang et al., 2024) refines speech-text alignment through knowledge distillation. SALMONN (Tang et al., 2023) aims to improve auditory comprehension of language and music in models. Qwen-Audio (Chu et al., 2023) advances audio recognition and translation by retraining speech encoders within a multi-task framework. Qwen2.5-Omni (Xu et al., 2025), the end-to-end multimodal model, is designed for comprehensive multimodal perception, seamlessly processing heterogeneous input modalities.

5 Conclusion

In this paper, we propose a novel strategy that reformulates the speech-to-text translation task as a combination of speech recognition and translation tasks, leveraging the machine translation capabilities of LLMs to enhance the performance of MLLMs in S2TT. To validate our approach, we train three MLLMs with sizes 3B, 7B, and 32B, and implement a three-stage curriculum learning strategy, which proves highly effective in low-resource scenarios while further improving performance when sufficient training data is available. Our model achieves state-of-the-art results across 15×14 translation directions, excelling in low-resource learning on the FLEURS dataset and supervised training on the CoVoST-2 dataset. These results highlight the robustness and effectiveness of our approach across diverse linguistic and data availability settings.

For future work, we aim to further optimize the LLM-SRT model to push its performance boundaries and extend its application to a broader range of languages.

Limitations

This paper presents a method for training an MLLM for languages with less than 10 hours of speech translation data.

However, the performance of S2TT and the range of supported languages are constrained by the capabilities of the LLM. MLLMs trained using this method may not perform well on languages that are not supported by the LLM or on those with poor machine translation performance.

Acknowledgement

The research in this article is supported by the National Natural Science Foundation of China (Grants No. U22B2059 and 62276083) and the National Science and Technology Innovation 2030 Major Program (Grant No. 2024ZD01NL00101). We also appreciate the support from China Mobile Group Heilongjiang Co., Ltd. @ on our research, the research is a joint collaboration between the involved parties.

References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Daniel Beck, Trevor Cohn, and Gholamreza Haffari. 2019. Neural speech translation using lattice transformations and graph networks. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 26–31.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019. Breaking the data barrier: Towards robust speech translation via adversarial stability training. *arXiv preprint arXiv:1909.11430*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Consecutive decoding for speech-to-text translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12738–12748.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. 2020. Espnet-st: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234*.

Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. 2023. Pre-training for speech translation: Ctc meets optimal

transport. In *International Conference on Machine Learning*, pages 18667–18685. PMLR.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. 2025. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv e-prints*, pages arXiv–2505.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

David Mueller, Mark Dredze, and Nicholas Andrews. 2024. Multi-task transfer matters during instruction-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14880–14891.

Sara Papi, Peidong Wang, Junkun Chen, Jian Xue, Naoyuki Kanda, Jinyu Li, and Yashesh Gaur. 2024. Leveraging timestamp information for serialized joint streaming recognition and translation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10381–10385. IEEE.

Trinh Pham, Khoi M Le, and Luu Anh Tuan. 2024. Unibridge: A unified approach to cross-lingual transfer learning for low-resource languages. *arXiv preprint arXiv:2406.09717*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421.

Matthias Sperber, Hendra Setiawan, Christian Gollan, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Consistent transcription and translation of speech. *Transactions of the Association for Computational Linguistics*, 8:695–709.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, et al. 2022. No language left behind: Scaling human-centered machine translation. Authors abbreviated for brevity. Full list at https://arxiv.org/abs/2207.04672.

Changhan Wang, Juan Pino, and Jiatao Gu. 2020a. Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation. *arXiv preprint arXiv:2006.05474*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Sravya Popuri, Dmytro Okhonko, and Juan Pino. 2020b. Fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Changhan Wang, Anne Wu, and Juan Pino. 2020c. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. 2024. Blsp-kd: Bootstrapping language-speech pre-training via knowledge distillation. *arXiv preprint arXiv:2405.19041*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Wenyi Yu, Changli Tang, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Connecting speech encoder and large language model for asr. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12637–12641. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

A Appendix

In this chapter, we report the language support in Table 10, the case study in Table 17, and the BLEU scores for the 15x14 directions of the FLEURS dataset in Table 18.

Code	Language	Family	Subgrouping	Script
deu	German	Indo-European	Germanic	Latn
eng	English	Indo-European	Germanic	Latn
fra	French	Indo-European	Italic	Latn
ind	Indonesian	Austronesian	Malayo-Polynesian	Latn
ita	Italian	Indo-European	Italic	Latn
jpn	Japanese	Japonic	Japanesic	Jpan
kor	Korean	Koreanic	Korean	Kore
nld	Dutch	Indo-European	Germanic	Latn
por	Portuguese	Indo-European	Italic	Latn
rus	Russian	Indo-European	Balto-Slavic	Cyrl
spa	Spanish	Indo-European	Italic	Latn
tha	Thai	Tai-Kadai	Kam-Tai	Thai
vie	Vietnamese	Austroasiatic	Vietic	Latn
yue	Cantonese	Sino-Tibetan	Sinitic	Hant
zho	Chinese	Sino-Tibetan	Sinitic	Hans

Table 10: **Language Support.** The table lists language codes, names, families, subgroups, and scripts. Language support is extensible based on ISO 639-3. 6×12 translation directions cover source languages (eng, deu, fra, zho, rus, jpn) and target languages (eng, deu, fra, spa, por, ita, nld, rus, jpn, kor, vie, ind, zho).

Stage	Dataset	Hour	Batch	Step	Learning Rate	Warmup Step	Optimizer
ASR	Common Voice 19	4498	16	472000	1e-4	1000	AdamW
	FLEURS	129			1e-4		
SMT	FLEURS	129		44000	1e-4		
SRT	FLEURS	129		83000	1e-5		

Table 11: **Training Details for LLM-SRT-3B-V2.** The step count refers to the number of steps on a single GPU.

Model	Encoder	Adapter	LLM	Language Support
LLM-SRT-3B	Whisper large-v3's encoder	Q-Former + MLP	Qwen2.5-3B	6 × 12
LLM-SRT-7B			Qwen2.5-7B	6 × 12
LLM-SRT-32B			Qwen2.5-32B	6 × 12
LLM-SRT-3B-V2			Qwen2.5-3B	15 × 14

Table 12: **Model Settings.** The V2 model employs the same architecture but supports more languages.

Model	Rank (r)	Alpha	Dropout	Target Keys	Bias
SeamlessM4T-V2	32	64	0.2	`.*_proj`	`none`
LLM-SRT-3B-V2	8	32	0.05	`q_proj, v_proj`	`none`

Table 13: **LoRA Configuration Settings.** We only fine-tuned the LLM with LoRA for LLM-SRT-3B-V2.

Modules	Param	Training stage	Details
Speech Encoder	∼635M	–	Whisper's encoder
Speech Adapter	∼73.7M	I&II&III	Q-Former and MLP
LLM	∼3.1B	–	Qwen2.5-3B
LLM adapter	∼1.8M	III	LoRA
Total	∼3.8B		

Table 14: **LLM-SRT-3B-V2 Training Parameters.** The red indicates trainable parameters.

X→14	SeamlessM4T-V2	SeamlessM4T-V2+Lora	LLM-SRT-3B	LLM-SRT-3B+Lora
deu	21.6 / 76.2	22.0 / 76.8	23.4 / 83.4	24.5 / 83.5
eng	33.3 / 84.8	32.8 / 84.3	27.8 / 85.5	29.1 / 85.4
fra	21.1 / 77.6	21.3 / 78.0	24.0 / 83.5	24.3 / 83.6
ind	18.4 / 73.8	19.1 / 75.0	22.4 / 82.3	22.9 / 82.2
ita	19.1 / 76.9	19.4 / 77.4	22.0 / 83.5	22.8 / 83.5
jpn	14.5 / 72.4	14.4 / 72.5	15.3 / 80.1	16.5 / 80.3
kor	17.2 / 76.6	17.0 / 76.5	17.9 / 81.8	18.0 / 81.9
nld	18.4 / 74.8	18.6 / 75.3	20.4 / 81.3	20.9 / 81.4
por	18.8 / 73.3	18.6 / 73.5	25.7 / 84.1	26.2 / 84.2
rus	20.6 / 75.9	21.3 / 76.7	22.8 / 82.5	23.3 / 82.5
spa	17.6 / 76.1	18.1 / 76.9	21.9 / 83.9	22.6 / 83.9
tha	12.8 / 71.1	13.2 / 71.7	12.5 / 78.6	14.7 / 78.8
vie	17.5 / 74.1	17.5 / 74.1	18.2 / 79.7	19.0 / 79.8
yue	15.2 / 70.9	14.9 / 70.5	15.5 / 78.9	16.3 / 79.0
zho	16.8 / 77.5	17.0 / 77.7	19.2 / 82.5	19.6 / 82.6
Avg.	18.8 / 75.5	19.0 / 75.8	20.6 / 82.1	21.4 / 82.2

Table 15: **BLEU/COMET Scores for 15 Languages Across 15×14 Translation Directions on FLEUR.** Detailed results are summarized in Tables 18.

Case		BLEU↑
Audio	There may be more maria on the near side because the crust is thinner. It was easier for lava to rise up to the surface. 地が薄いため、近い方には海が多くなることがあります。溶岩が浮上しやすくなっていました。	
Qwen2.5-3B (MT)	近には可能にマリアがあるかもしれませんbecause the crustははがより薄いからマagmaが表面に上りやすかったlavaです。	2.9
LLM-SRT-3B (S2TT)	近にはもっとマリアがある可能性があります。地が薄いためです。溶岩が表面に上りやすかったからです。	30.6

Table 16: **Error Analysis.** We observe that the Qwen2.5-3B model exhibits language mixing (e.g., Japanese and Cantonese) in certain translation directions. Consequently, S2TT outperforms MT in these scenarios.

Case		BLEU↑
Audio # 1	它让玩家可以通过在空中移动设备来控制电子游戏中的运动和操作。 This will allow players to control actions and movements in video games by moving the device through the air.	
SeamlessM4T-V2	It allows players to control the movement and operation of electronic games through mobile devices in the air.	14.0
LLM-SRT-3B	它让玩家可以通过在空中移动设备来控制电子游戏中的动作和操作。<|zho|><|eng|> It allows players to control the actions and operations of an electronic game by moving the device in the air.	23.2
LLM-SRT-32B	它让玩家可以通过在空中移动设备来控制电子游戏中的动作和操作。<|zho|><|eng|> It allows players to control actions and operations in electronic games by moving the device in the air.	43.0
Audio # 2	In fact, it is not easy to find at all even if one knew it existed. Once inside the cave, it is a total isolation. 事实上，即使知道它的存在，也不容易找到。一旦进入洞穴，就完全与世隔绝了。	
SeamlessM4T-V2	事实上,即使有人知道它存在, 也很难找到它. 一旦进入洞穴,	22.7
LLM-SRT-3B	In fact, it is not easy to find it at all, even if one knew it existed. Once inside the cave, it is a total isolation.<|eng|><|zho|> 事实上，即使知道它存在，要找到它也是很困难的。一旦进入洞穴，就是完全的隔离。	49.3
LLM-SRT-32B	In fact, it is not easy to find at all, even if one knew it existed. Once inside the cave, it is a total isolation. <|eng|><|zho|> 事实上，即使知道它的存在，也很难找到。一旦进入洞穴，就完全与世隔绝了。	86.8

Table 17: **Case Study.** We compare the BLEU scores of our models with SeamlessM4T-V2.

src_tgt	SeamlessM4T-V2 BLEU	Avg.	+Lora	Avg.	LLM-SRT-3B-V2 BLEU	Avg.	+Lora	Avg.
deu_eng	37.1		36.9		32.3		34.3	
deu_fra	19.3		21.2		23.2		25.0	
deu_ind	19.1		19.4		17.2		18.8	
deu_ita	12.5		13.5		13.8		14.8	
deu_jpn	26.9		25.8		30.2		31.9	
deu_kor	29.1		29.0		23.1		24.6	
deu_nld	13.2	21.6	13.8	22.0	14.2	23.4	15.0	24.5
deu_por	19.7		21.4		22.2		23.5	
deu_rus	13.3		14.7		14.5		15.9	
deu_spa	13.7		14.8		16.2		17.2	
deu_tha	44.3		41.9		43.6		45.2	
deu_vie	22.0		22.0		21.3		23.0	
deu_yue	12.2		12.8		24.9		25.5	
deu_zho	20.5		21.1		30.8		28.5	
eng_deu	33.4		33.1		21.2		22.7	
eng_fra	42.6		41.8		33.0		34.7	
eng_ind	37.1		37.2		25.6		27.4	
eng_ita	24.5		24.6		18.3		19.2	
eng_jpn	36.0		34.2		33.7		37.3	
eng_kor	40.2		38.7		26.4		28.3	
eng_nld	24.4	33.3	24.2	32.8	15.5	27.8	15.9	29.1
eng_por	42.9		42.9		34.5		36.1	
eng_rus	26.2		25.8		19.4		20.2	
eng_spa	23.7		24.0		20.7		22.1	
eng_tha	52.5		50.5		46.3		50.0	
eng_vie	36.1		35.9		27.6		29.7	
eng_yue	16.6		17.2		29.9		30.5	
eng_zho	30.1		28.7		37.0		34.1	
fra_deu	10.6		12.0		16.0		16.4	
fra_eng	33.8		33.7		32.8		34.7	
fra_ind	19.0		20.1		18.8		19.8	
fra_ita	15.3		15.9		16.9		17.1	
fra_jpn	27.9		26.6		30.6		31.1	
fra_kor	29.1		28.9		23.1		23.4	
fra_nld	11.0	21.1	12.0	21.3	12.7	24.0	12.8	24.3
fra_por	20.0		20.9		24.5		25.5	
fra_rus	12.2		12.9		15.7		15.9	
fra_spa	14.3		14.6		18.5		18.9	
fra_tha	46.2		45.1		45.8		45.7	
fra_vie	22.4		23.1		22.9		24.3	
fra_yue	13.2		12.8		27.5		25.4	
fra_zho	20.1		20.2		30.7		29.2	
ind_deu	10.3		12.5		13.3		14.1	
ind_eng	32.4		32.5		30.5		32.8	
ind_fra	16.2		17.4		21.5		22.4	
ind_ita	9.8		11.0		13.1		13.7	
ind_jpn	24.8		23.4		29.7		30.6	
ind_kor	25.9		25.5		22.0		23.2	
ind_nld	10.0	18.4	10.6	19.1	10.8	22.4	11.2	22.9
ind_por	17.0		19.3		20.3		21.5	
ind_rus	8.2		9.8		13.8		13.9	
ind_spa	10.8		11.7		14.5		15.5	
ind_tha	43.9		44.6		45.5		46.2	
ind_vie	19.3		19.7		22.9		23.4	
ind_yue	11.7		11.7		25.4		25.7	
ind_zho	17.4		17.9		29.8		26.7	
ita_deu	10.5		11.6		13.4		14.3	
ita_eng	26.5		26.1		24.7		27.1	
ita_fra	17.6		18.5		22.3		23.9	
ita_ind	15.3		16.3		15.3		16.6	
ita_jpn	24.7		23.7		29.8		30.6	
ita_kor	26.8		26.6		22.0		23.3	
ita_nld	9.6	19.1	10.3	19.4	10.2	22.0	10.6	22.8
ita_por	15.4		15.9		19.4		21.4	
ita_rus	11.4		12.1		13.0		13.9	
ita_spa	13.9		14.2		16.3		17.7	
ita_tha	44.7		43.2		45.1		45.6	
ita_vie	19.1		20.0		22.1		23.0	
ita_yue	12.8		12.8		25.3		24.4	
ita_zho	19.1		20.4		29.3		26.2	
jpn_deu	8.5		8.6		7.9		9.1	
jpn_eng	18.2		17.8		16.5		19.3	
jpn_fra	14.7		14.5		13.4		15.0	
jpn_ind	11.4		12.6		10.3		11.7	
jpn_ita	9.0		8.9		8.5		9.2	
jpn_kor	25.5		25.8		24.5		26.0	
jpn_nld	9.3	14.5	9.3	14.4	7.8	15.3	8.3	16.5
jpn_por	13.4		13.7		13.4		14.9	
jpn_rus	7.7		7.6		7.7		8.6	
jpn_spa	11.6		11.7		10.9		11.8	
jpn_tha	37.0		34.0		38.4		40.2	
jpn_vie	15.8		15.6		14.9		16.4	
jpn_yue	7.7		8.2		19.1		19.4	
jpn_zho	13.2		13.5		21.7		21.2	
kor_deu	9.4		9.3		8.7		9.1	
kor_eng	24.3		22.9		20.3		21.9	
kor_fra	16.5		16.4		15.0		14.9	
kor_ind	16.7		16.2		13.3		13.5	
kor_ita	10.3		10.3		9.2		9.9	
kor_jpn	28.3		28.4		31.5		32.1	
kor_nld	8.9	17.2	9.6	17.0	8.1	17.9	8.3	18.0
kor_por	15.3		15.6		15.7		16.1	
kor_rus	9.8		10.0		9.1		8.9	
kor_spa	12.7		12.8		12.4		12.9	
kor_tha	39.2		36.0		41.2		41.1	
kor_vie	19.2		19.4		18.5		19.3	
kor_yue	12.3		13.1		23.2		21.5	
kor_zho	17.9		18.0		25.0		22.4	
nld_deu	11.4		11.7		12.3		12.7	
nld_eng	27.8		27.4		23.6		25.7	
nld_fra	15.2		15.9		19.4		19.7	
nld_ind	15.6		15.9		14.7		15.1	
nld_ita	10.0		10.7		10.9		11.9	
nld_jpn	24.3		23.4		27.3		29.0	
nld_kor	24.7	18.4	25.1	18.6	21.3	20.4	22.7	20.9
nld_por	15.2		15.9		16.6		18.3	
nld_rus	10.3		11.5		12.2		12.5	
nld_spa	12.8		13.3		13.6		14.8	
nld_tha	41.8		40.5		43.6		43.7	
nld_vie	17.8		18.0		18.9		20.2	
nld_yue	11.4		11.7		23.3		23.9	
nld_zho	19.2		19.4		27.3		23.3	

src_tgt	SeamlessM4T-V2 BLEU	Avg.	+Lora	Avg.	LLM-SRT-3B-V2 BLEU	Avg.	+Lora	Avg.
por_deu	6.8		7.5		18.2		18.5	
por_eng	38.4		37.7		38.3		40.8	
por_fra	16.3		16.7		28.8		29.6	
por_ind	13.9		15.2		20.5		21.6	
por_ita	10.3		10.4		17.4		17.5	
por_jpn	24.5		23.4		31.8		32.8	
por_kor	27.8	18.8	26.0	18.6	23.8	25.7	24.9	26.2
por_nld	9.8		10.0		12.8		13.3	
por_rus	9.1		9.3		17.2		18.0	
por_spa	13.2		13.1		18.5		20.1	
por_tha	44.2		42.7		46.6		47.1	
por_vie	19.0		19.2		25.2		25.9	
por_yue	11.6		11.8		28.3		27.2	
por_zho	17.9		17.9		32.7		30.1	
rus_deu	15.2		16.3		14.9		15.5	
rus_eng	30.2		29.6		28.7		30.0	
rus_fra	20.8		22.5		22.2		23.3	
rus_ind	17.9		19.0		16.9		18.1	
rus_ita	14.7		15.7		14.0		14.0	
rus_jpn	22.8		23.0		29.5		30.7	
rus_kor	25.6	20.6	25.0	21.3	21.5	22.8	23.1	23.3
rus_nld	13.0		13.9		11.6		11.6	
rus_por	20.4		21.6		22.1		22.9	
rus_spa	15.3		16.0		16.4		16.6	
rus_tha	43.1		41.6		44.7		44.4	
rus_vie	21.9		23.1		21.8		23.1	
rus_yue	10.4		11.9		24.4		24.6	
rus_zho	16.9		18.7		29.8		28.7	
spa_deu	7.9		10.1		12.8		13.4	
spa_eng	25.4		25.4		24.5		27.8	
spa_fra	13.5		14.7		22.0		23.2	
spa_ind	12.9		14.7		16.0		17.3	
spa_ita	11.1		11.8		14.8		15.9	
spa_jpn	24.4		23.4		29.6		30.2	
spa_kor	27.6	17.6	26.5	18.1	21.8	21.9	22.7	22.6
spa_nld	9.2		10.2		10.3		11.6	
spa_por	14.4		15.0		19.1		21.0	
spa_rus	8.3		9.8		13.7		14.1	
spa_tha	44.6		43.9		45.0		45.8	
spa_vie	17.4		18.8		21.8		23.1	
spa_yue	12.0		11.9		25.2		24.4	
spa_zho	17.4		17.8		29.7		26.3	
tha_deu	9.3		9.6		7.8		9.7	
tha_eng	23.3		22.7		14.4		19.1	
tha_fra	13.4		14.1		11.8		14.1	
tha_ind	11.6		13.0		10.3		13.2	
tha_ita	8.6		9.5		7.2		8.7	
tha_jpn	17.2		16.3		22.6		23.9	
tha_kor	20.8	12.8	20.1	13.2	15.4	12.5	19.2	14.7
tha_nld	8.4		9.1		6.8		8.2	
tha_por	12.9		13.9		10.8		14.0	
tha_rus	9.0		10.0		7.9		9.4	
tha_spa	10.0		10.7		9.8		11.6	
tha_vie	14.3		15.2		14.6		17.3	
tha_yue	8.0		8.0		16.1		18.1	
tha_zho	12.4		12.8		19.6		20.1	
vie_deu	11.5		11.5		10.4		11.3	
vie_eng	25.6		25.2		21.1		23.8	
vie_fra	17.5		17.8		16.6		17.3	
vie_ind	15.9		17.5		14.2		15.7	
vie_ita	11.3		11.5		10.2		11.2	
vie_jpn	21.8		20.6		25.0		26.2	
vie_kor	23.9	17.5	21.7	17.5	19.9	18.2	20.4	19.0
vie_nld	10.6		10.9		8.5		9.3	
vie_por	16.4		16.8		15.6		17.0	
vie_rus	12.1		12.1		11.0		11.6	
vie_spa	12.1		12.3		12.7		13.8	
vie_tha	40.6		39.4		42.6		43.1	
vie_yue	11.4		12.0		21.8		21.0	
vie_zho	14.8		15.1		25.0		24.1	
yue_deu	8.6		8.6		7.6		8.2	
yue_eng	18.8		17.8		16.1		18.2	
yue_fra	13.4		13.5		13.6		14.0	
yue_ind	12.2		12.9		10.4		11.8	
yue_ita	10.0		9.8		8.9		9.3	
yue_jpn	14.8		14.8		22.9		24.6	
yue_kor	19.9	15.2	19.4	14.9	18.2	15.5	20.5	16.3
yue_nld	8.0		7.9		7.1		7.5	
yue_por	12.9		12.9		12.1		12.8	
yue_rus	8.6		8.4		8.2		8.4	
yue_spa	10.8		10.9		10.8		11.1	
yue_tha	35.2		32.8		38.7		39.9	
yue_vie	15.7		16.0		16.0		17.1	
yue_zho	23.4		23.3		26.7		24.8	
zho_deu	10.2		10.4		9.6		11.0	
zho_eng	21.7		22.4		23.1		23.9	
zho_fra	16.7		16.6		17.4		18.2	
zho_ind	13.9		15.5		14.0		14.8	
zho_ita	11.6		11.4		10.7		11.3	
zho_jpn	18.1		17.9		27.1		28.8	
zho_kor	23.4	16.7	23.4	17.0	22.3	19.2	23.0	19.6
zho_nld	10.8		10.9		8.7		8.9	
zho_por	16.2		16.4		16.7		17.3	
zho_rus	9.8		10.6		10.1		10.8	
zho_spa	12.6		13.2		14.2		14.6	
zho_tha	38.8		38.0		42.0		43.0	
zho_vie	18.0		18.7		20.0		20.6	
zho_yue	12.3		12.9		32.5		28.7	
Avg.	18.8		19.0		20.6		21.4	

Table 18: **BLEU Scores for 15 Languages Across 15×14 Translation Directions on FLEURS.**